Superintelligence and the Future of Governance: On Prioritizing the Control Problem at the End of History

Phil Torres
Published in Roman Yampolskiy (ed.), *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC Press.

Abstract: This chapter argues that dual-use emerging technologies are distributing unprecedented offensive capabilities to nonstate actors. To counteract this trend, some scholars have proposed that states become a little "less liberal" by implementing large-scale surveillance policies to monitor the actions of citizens. This is problematic, though, because the distribution of offensive capabilities is also undermining states' capacity to enforce the rule of law. I will suggest that the only plausible escape from this conundrum, at least from our present vantage point, is the creation of a "supersingleton" run by a friendly superintelligence, founded upon a "post-singularity social contract." In making this argument, the present chapter offers a novel reason for prioritizing the "control problem," i.e., the problem of ensuring that a greater-than-human-level AI will positively enhance human well-being.

1. Introduction

Several theorists in the past few centuries have declared that history has reached or someday will reach a *telos*, at which point the evolution of sociocultural institutions or political systems effectively cease. For example, Georg Wilhelm Friedrich Hegel claimed that history came to an end in 1806 with the battle of Jena and Karl Marx posited that the teleological terminus of history would be a world communist society. More than a century after Marx, Francis Fukuyama published a 1989 article, later expanded into *The End of History and the Last Man* (1992), in which he argues that the collapse of the Soviet Union marks not just the conclusion of a post-war episode of international tensions, "but the end of history as such: that is, the end point of mankind's ideological evolution and the universalization of Western liberal democracy as the final form of human government" (Fukuyama 1992). Hegel, Marx, and Fukuyama were, of course, wrong about their grand narrative proclamations, and indeed Fukuyama later admitted that

> in the course of thinking through the many critiques of that original piece that had been put forward [in 1989], it seemed to me that the only one that was not possible to refute was the argument that there could be no end of history unless there was an end of science. As I had described the mechanism of a progressive universal history in … [*The*] *End of History and the Last Man*, the unfolding of modern natural science and the technology that it spawns emerges as one of its chief drivers. Much of late-twentieth-century technology, like the so-called Information Revolution, was quite conducive to the spread of liberal democracy. But we are nowhere near the end of science.

Here I propose a provocative question: Or are we? That is, are there reasons for thinking that the end of science could be "near" on timescales meaningful to contemporary humans? Consider the following claims: first, according to recent surveys, humanity could create a human-level artificial general intelligence by 2100 (Müller and Bostrom 2014). Second, if we create a human-level artificial general intelligence through an "extendible" method, it could be followed shortly after by a superintelligence, meaning that there is a good chance that we will create a superintelligence by 2100 (Chalmers 2010). Third, if this occurs, it will likely mark an end to the *human* scientific enterprise for the very same reason that, to quote I.J. Good (1964), "the first ultraintelligent machine is the last invention that man need ever make." But it could also mark the end of science *as such*, since acquiring a complete explanatory-predictive "theory of everything" would likely constitute an instrumental value for any given superintelligence with any set of final goals. It follows that, if the critique that Fukuyama singles-out above is correct, the first superintelligence could mark the end of history in a very significant sense.[1]

The present article will argue that we may need an "end to history" in the form of a *friendly supersingleton* to overcome the immense dangers posed by the democratization of science and technology.[2] By "friendly supersingleton" I mean a singleton, or global governing system, that is run by a friendly superintelligence, or a generally intelligent algorithm that (a) far exceeds the performance of human brains in every cognitive domain, and (b) has a value system that makes its behavior conducive to human flourishing. The implementation of a friendly supersingleton in the relative near-term will almost certainly be *sufficient* but could also be *necessary* for humanity to avoid an existential catastrophe in the coming decades or centuries. For the sake of clarity, the premises of my argument are as follows:

> (i) *The Threat of Universal Unilateralism*: Emerging technologies are enabling a rapidly growing number of nonstate actors to unilaterally inflict unprecedented harm on the global village; this trend of mass empowerment is significantly increasing the probability of an existential catastrophe—and could even constitute a Great Filter (Sotos 2017).
> (ii) *The Preemption Principle*: If we wish to obviate an existential catastrophe, then societies will need a way to preemptively avert not just most but all possible attacks with existential consequences, since the consequences of an existential catastrophe are by definition irreversible.[3]
> (iii) *The Need for a Singleton*: The most effective way to preemptively avert attacks is through some regime of mass surveillance that enables governing bodies to monitor the actions, and perhaps even the brain states, of citizens; ultimately, this will require the formation of a singleton.

---

[1] Even more to the point, Fukuyama (2002) worries that person-engineering technologies could alter our human nature, which is what, he claims, liberal democracy is founded upon; thus, person-engineering technologies could undermine liberal democracy. Yet the proposal here advanced replaces liberal democracy with a mixed democratic/autocratic friendly supersingleton whose legitimacy would not be undermined if morphological freedom were the laws of the land and our evolutionary lineage were to rapidly diversity into a wide variety of posthuman forms.
[2] Here I am ignoring the other two "Great Challenges" of our time, namely, anthropogenic climate change and nuclear proliferation. See Torres 2017e.
[3] Given the distinction between terror and error, this *should* read "attacks or accidents." For the present discussion, I am focusing primarily on agential terror.

(iv) *The Threat of State Dissolution*: The trend of (i) will severely undercut the capacity of governing bodies to effectively monitor their citizens, because the capacity of states to provide security depends upon a sufficiently large "power differential" between themselves and their citizens.[4]
(v) *The Limits of Security*: If states are unable to effectively monitor their citizens, they will be unable to neutralize the threat posed by (i), thus resulting in a high probability of an existential catastrophe.[5]

The following sections will attempt to justify each of these premises. Then, in the penultimate section, I will present an argument for why a superintelligence designed to govern human affairs could avoid the dangerous outcome of (v). The moral of this story will be that humanity must solve the "control problem" in the field of AI risk—and solve it *soon*—not merely because a value-misaligned superintelligence might convert humanity into paperclips (see Bostrom 2014), but because we may need a value-aligned superintelligence to overcome the "Great Challenge" of more powerful and accessible technologies (see Torres 2017e). Even more, we must solve this problem *before* the time at which distributed offensive capabilities begin to threaten the modern state system (as illustrated in Figure 2). This paper thus offers a new reason for prioritizing AI safety research, focusing in particular on the design of greater-than-human-level general intelligence algorithms capable of implementing top-down policies that can ensure our collective survival and prosperity.

2. The Growing Capacity for Unilateral Destruction

Who among us would destroy the world?[6] In the personal journal of Eric Harris, the psychopathic mastermind behind the 1999 Columbine High School massacre, one finds the following sentence: "I think I would want us to go extinct. … I just wish I could actually DO this instead of just DREAM about it all" (quoted in Langman 2010). This shocking statement expresses one of two necessary conditions for token "agential risks," as I have elsewhere called them, to realize their omnicidal fantasies, namely, the *motivation* (Torres 2017a, 2017b, 2017c, 2017d). The other necessary condition is of course the *means*. Together these are sufficient for such individuals to cause harm.

Fortunately, our technological civilization has not yet reached the stage at which the means are readily available to deranged agents with a death wish for humanity.[7] Yet emerging technologies are changing this situation fast, at an exponential or even super-exponential pace. The reason concerns three crucial features of emerging technology, namely, their (a) dual usability, (b) power, and (c) accessibility (see Torres 2017a). Briefly put, a technology is dual-use if and

---

[4] For more on this crucial point, see Wittes and Blum 2015; it is their discussion of the topic that has, in part, inspired the present paper.
[5] Thanks to Matthijs Maas for suggesting the appellations given to each premise.
[6] This is the central question of agential risk studies. See Torres 2017b.
[7] An agential risk refers to any agent who could pose a threat to humanity or human civilization if she or he were to gain access to a doomsday button, where a doomsday button would, if pressed, initiate a "weapon of total destruction," or WTD (Torres 2017a).

only if it can be employed for both morally good and bad ends.[8] This may sound trivial since *all* human-made artifacts could, given the inevitable malleability of even the most specialized technical designs, be used for both such ends. For example, someone could weaponize a laptop by using it as a bludgeon to beat another person to death. Nonetheless, dual usability becomes exceedingly relevant when the potential bad uses could seriously damage civilization, or perhaps even initiate an existential disaster. Seth Baum refers to this as the "great downside dilemma" of advanced technologies (Baum 2014).

Second, advanced technologies are enabling users to manipulate and rearrange the physical world in increasingly significant ways (see Figure 1). The most obvious historical example of a new technology introducing a sudden discontinuity in human destructive capabilities is nuclear weapons. These can initiate massive firestorms in urban areas that pollute the stratosphere with sunlight-blocking soot, thereby causing a nuclear winter that devastates global agriculture and drastically reduces the human population (see, e.g., Roebuck et al. 2007). But there are other types of emerging technologies that could produce similarly catastrophic outcomes, including biotechnology, synthetic biology, molecular nanotechnology, and "tool AI."[9] Consider that in 2001 scientists demonstrated (by accident, as it happens) that genetic engineering can greatly increase the virulence of pathogens. This affirms that malicious agents could synthesize designer germs that are far more dangerous than anything cooked up in the Darwinian laboratory of nature. At the extreme, it is theoretically possible to create a germ that combines the lethality of rabies, the incurability of Ebola, the contagiousness of the common cold, and the long incubation period of HIV (see Torres 2017a). The result could be an accidental or intentional release that brings about a global pandemic worse than anything humanity (or any other species) has ever before experienced. Atomically precise manufacturing—or the manipulation of matter with

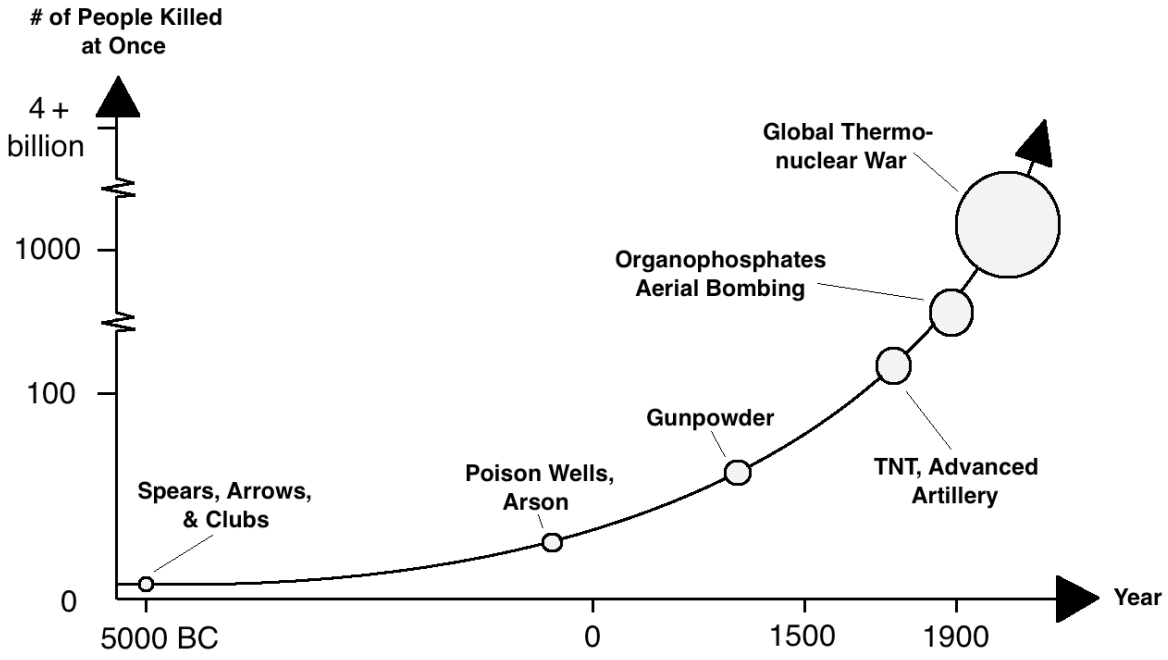

Figure 1*: Diagram illustrating the growing power of technology. Note that this does not take into account the increasingly destructive capabilities of biotechnology, synthetic biology, molecular nanotechnology, and "tool AI." (Based on a figure created by Gary Ackerman; see Torres 2017a.)*

absolute atomic precision—could have even more devastating effects (see Drexler 2013). For example, someone could design a self-replicating autonomous nanobot that converts all the organic matter it comes into contact with into clones of itself, thereby transmogrifying the entire biosphere into a wriggling swarm of mindlessly reproducing machines.

Third, the unprecedented power of advanced technologies is being "democratized"—i.e., placed within reach of a growing number of agents. There are four axes along which this trend is unfolding. First, the *intelligence threshold* that one must exceed to bring about large-scale destruction is dropping. Eliezer Yudkowsky (2008) captures this idea with his "Moore's Law of Mad Science," which states that "every eighteen months, the minimum IQ necessary to destroy the world drops by one point" (see Torres 2017a). Second, the *information threshold* that one must exceed to use advanced technologies competently is also dropping. Many laboratory processes, for example, can be followed like a recipe for chocolate cake, and the full genomes of bugs like smallpox and Ebola are publicly available online. Third, the *skill threshold* that one must exceed to competently turn one's "know-that" into "know-how" is dropping as well. This is most salient with synthetic biology, which is "*explicitly devoted* to the minimization of the importance of tacit knowledge" (Mukunda et al. 2009). The BioBricks Foundation's approach to standardizing biological entities and digital-to-biological converters are especially relevant here (see Boles et al. 2017). But it could become even more significant with respect to molecular nanotechnology. And fourth, the *materials or equipment* needed to wield existentially dangerous technologies are becoming cheaper and more widely available. Consider that nanofactories could manufacture other nanofactories and whereas highly enriched uranium needed for nuclear weapons has historically been difficult to obtain, new laser enrichment technologies (e.g., SILEX) could enable small groups or individuals to produce this key ingredient.

The result of these three macro-trends and the micro-trends that they subsume is what we can call the "threat of universal unilateralism."

> *Threat of universal unilateralism*: emerging technologies are distributing unprecedentedly destructive offensive capabilities to both state and nonstate actors; in doing so, they are rapidly multiplying the total number of agents capable of inducing a global or even existential disaster.

A society in which a large number of individuals have access to a "doomsday button," as it were, would find itself in a frightfully precarious existential predicament. To underline just how dangerous this situation would be, consider some recent calculations by John Sotos. Focusing entirely on dual-use artifacts within the biological sciences, he finds that a 1 in 100 chance of only a few hundred agents releasing a species-destroying pathogen yields virtually inevitable doom within ~100 years. Even more, if the total number of agents capable of inflicting existential harm rises to 100,000, the probability of *someone* releasing such a pathogen must be less than 1 in 10^9

for civilization to survive a millennium (MIT 2017). This leads Sotos to conclude that if "civilizations universally develop advanced biology, before they become vigorous interstellar colonizers, [then this] model provides a resolution to the Fermi paradox" (Sotos 2017). I have elsewhere (and independently) proposed similar calculations based on my own models of agential risk: let us posit just 1,000 terror agents in a population of 10 billion and that the probability per decade of any one of these individuals gaining access to world-destroying weapons is only 1 percent. What overall level of existential risk would this expose the entire population to? It turns out that, given these assumptions, the probability of doom per decade would be a staggering 99.995 percent. One gets the same result if the number of terror agents is 10,000 and the probability of access is 0.1 percent, or if the number is 10 million and the probability is 0.000001. Now consider that this probability may become *much greater* than 0.000001, or even 1 percent, and that the number of terror agents could plausibly reach 10 million, which is a mere 0.1 percent of 10 billion.[10] It appears that an existential catastrophe could be more or less *inescapable* given that, as Bostrom puts it, "some little idiot is bound to press the ignite button" (Bostrom 2014).[11]

The rest of this paper will assume that current techno-developmental trends continue such that the threat of universal unilateralism grows. Indeed, my own considered opinion is that technologization has become a juggernaut-like process without any breaks, i.e., technological development is largely "autonomous" from human control in certain crucial senses (Winner 1977). By analogy, just as the flight direction of a flock of starlings at dusk depends upon the individual actions of each bird in the murmuration, so too does technological development depend upon the actions of individuals, institutes, organizations, corporations, and governments—yet no single entity or entity-ensemble can manage the macro-trajectory of this evolutionary phenomenon. This is also consistent with Ray Kurzweil's (2005) assertion that the genetics, nanotech, and robotics (GNR) revolution is "inevitable" (bracketing major "defeaters" like an existential catastrophe), as well as Bostrom's (2009) "technological completion conjecture," which states that "if scientific and technological development efforts do not effectively cease, then all important basic capabilities that could be obtained through some possible technology will be obtained."[12] The point is that future civilization will, *ceteris paribus*, almost certainly witness the asymptotic realization of a condition of universal unilateralism and with it a global threat environment in which virtually everyone could pose an existential danger to humanity.

---

[10] This paragraph quotes from Torres 2017e.

[11] Wittes and Blum (2015) describe this general situation as "the outlying, extreme case." My own considered view is that we should see it as the *default outcome* of continued emerging tech development, which there is no reason to believe is going to stop (in the absence of a major catastrophe). Thus, the present paper addresses precisely the worry that Wittes and Blum outline in this passage, which is also quoted in section 5: "If the outlying, extreme case really comes to pass—if technologies of mass empowerment enable many isolated individuals or diverse nonstate actors to injure or violate other individuals on a mass scale anywhere around the globe with substantially reduced fear of detection and punishment—we are in big trouble. Much of civil and political life as we know it will likely come to an end."

[12] To be clear, there are a range of "autonomous technology" theses. At the extreme, some scholars are guilty of reifying technology as a kind of self-generating organism with its own in-built *telos*. My position is simply this: if *we* don't develop any given artifact X, then *someone else will*. This seems to hold for all the emerging technologies, and it is in this sense that the development of technology is beyond our control.

3. The Need for Global Surveillance

Perhaps the most obvious strategy for obviating a bad outcome is mass surveillance. In fact, this is precisely what Ingmar Persson and Julian Savulescu (2012) endorse in a broader discussion of agent-oriented risk mitigation strategies (see Torres, 2017b).[13] As they put it, society must become a little "less liberal" to avoid the terminal nightmare of "Ultimate Harm," i.e., an existential disaster that would "render worthwhile life forever impossible" (Persson and Savulescu 2012). Surveillance could employ a variety of mechanisms, both already in use and merely anticipated. For example, governments could exploit the global telecommunications system, such as cellular networks and the Internet, which now form the information circulatory system of a vast cybernetic organism upon which contemporary civilization depends for survival. Many governments are, of course, doing precisely this, enabled by legislation like the USA Patriot Act in the United States.

More speculatively, Benjamin Wittes and Gabriella Blum (2015) describe a nimble "surveillance spider" that could be remote-controlled, semi-autonomous, or autonomous. Due to its miniature size, this robotic spider could infiltrate buildings, rooms, vehicles, and other enclosed spaces, record audiovisual data, and then relay this data back to a control center for analysis. A similar possibility would involve not terrestrial "insects" but aerial creatures like robotic flies that could navigate the world via all three dimensions of space. Even more, some visionaries have imagined autonomous nanobots being transmitted like pollen stuck to one's shoes, or capable of moving themselves around in their environments. This "smart dust" could enable one to achieve the ultimate stealthy spying—dedicated machines just a few billionths of a meter in size collecting information and then transmitting it wirelessly to the relevant parties. Since nanobots tasked with defending the planet against gray goo have been called "blue goo," let's call nanobots designed for clandestine operations "dark blue goo."

Another set of options stems from mind-reading technologies. In recent years, this field of research has seen numerous major breakthroughs in both theoretical knowledge and practical application—steps forward that suggest the capacity to read the thoughts of others could someday become ubiquitous. For example, scientists have used brain waves to reconstruct movie clips being watched in realtime. A couple of scientists in 2014 managed to transmit messages to each other using a noninvasive brain-to-brain connection. NASA and collaborators are working on technology that would measure the brainwaves of someone driving a car and alert them if they become too sleepy. Researchers have created a decoder that can read the words of one's private internal monologue. And scientists have figured out a way to verify the identity of people entirely based on neurological responses to particular words. Even more, studies show that our brains decide between options upwards of ten seconds before we are conscious of the decision, meaning that scientists can predict our choices before we make them (Soon et al. 2008; see also Smith 2008).[14]

Although the use of dark blue goo or mind-reading systems to monitor the actions of a population may be anathema to contemporary norms, it could be the case that a series of non-ex-

---

[13] I.e., moral bioenhancement using *mostropics* (i.e., pharmaceutical moral enhancers) coupled with cognitive enhancers.

[14] Although see also Miller and Schwarz 2014.

istential global catastrophes this century pushes society toward accepting a major loss of individual privacy as the necessary cost of security. Imagine the public response to an act of nuclear terrorism that turns Manhattan, Tokyo, and London into smoldering graveyards. Even more, there are reasons for expecting two or more catastrophes occurring in succession, a phenomenon that I call *catastrophe clustering* (Torres 2017a). For example, Seth Baum and his colleagues outline a "double catastrophe" scenario in which a war, economic recession, or terrorist attack—the first catastrophe—interrupts an ongoing stratospheric geoengineering program, thereby causing a sudden and perhaps unsurvivable rise in global surface temperatures—the second catastrophe (Baum et al. 2013). Some omnicidal maniacs might also use an initial catastrophe as a springboard to initiate a second catastrophe whose consequences interact synergistically with the first, and indeed many natural/anthropogenic catastrophe scenarios occur at random, making them susceptible to the "clustering illusion" phenomenon that accurately models the onset (and termination) of wars throughout human history (Torres 2017a, 2017b; see also Pinker 2011).

It is also worth noting here that our modern sense of privacy is just that: *modern*. Only a few centuries ago our ancestors were accustomed to urinating, defecating, and even copulating in public spaces. Privacy was not recognized as the fundamental right that we see it as today. Indeed, Persson and Savulescu argue that privacy itself isn't a moral right the way life and liberty may be, although the *means* employed to acquire personal information and the *uses* to which this information are applied could violate one's rights (Persson and Savulescu 2012). It follows that one cannot infringe upon another's right to privacy, and this helps to justify pushing society toward more "illiberal" modes through the use of invasive surveillance techniques. Finally, we should note that there are instances in which some loss of privacy is conducive to the flourishing of liberty—that is, the relationship between the two is not straightforwardly linear. For example, the creation of a searchable database of sex offenders requires some individuals' private information being made public, yet many would contend that the net benefit to society is positive (see Wittes and Blum 2015).

There could also arise a culture of widespread sousveillance, or the use of wearable recording devices so that the surveillees (the citizens being watched) can surveil the surveillers (the government agents doing the watching) (Mann et al. 2003). This has already happened to some degree around the world, spurring the rise of social justice movements like Black Lives Matter in the wake of numerous unjustified murders of unarmed black people by police. At the very extreme, one could imagine a completely "transparent society" in which everyone can see what everyone else is doing all the time (Brin 1996). Unfortunately, this vision of reciprocal accountability appears to be unpromising. While transparency would enhance the capacity of law enforcement to track the activities and movements of token agential risks, it would also enable such agents to track the activities and movements of law enforcement, thus giving offenders a potential first-mover advantage amidst the chaos of the "real world." This is worrisome because as Bostrom (2002) points out, when it comes to omnicidal agents and existential risks, our approach must be entirely proactive rather than reactive, since an existential catastrophe can only happen *once* in a species' career. Yet the notion of accountability is backward-looking, so it's unclear how it could accomplish the forward-looking goal of preventing through preemptive action

even a single attack with existential implications from occurring.[15] (One may be reminded here of the Provisional Irish Republican Army that, after nearly assassinating Margaret Thatcher, declared that "today we were unlucky, but remember we only have to be lucky once. You will have to be lucky always.")

It thus appears that for a system of mass surveillance to be effective it will have to be unidirectionally transparent, or *asymmetrical*, enabling the state to watch its citizens but not *vice versa*. Furthermore, given the growing capacity of nonstate actors to cause mass destruction, this system will also need to be *invasive*, instantiating something like the Ultimate Panopticon from which the relevant agencies can observe all the going-ons of all of society's members all the time.[16] But empowering the multiplicity of states within the international arena to monitor their citizens might not be enough in a world where individuals have the unilateral capacity to harm nearly anyone (or everyone) else on the planet. As Wittes and Blum (2015) observe,

> if nonstate actors can routinely challenge the authority of even strong states from within their territories, as well as from outside their territories, can the state still effectively serve a primary security function? … Do we not need some form of world government if we are effectively to police a globe in which anyone anywhere can attack anyone else anywhere else?

The idea of a global governing system has been imagined in various forms by thinkers for centuries. Dante considered it within the Christian context, Immanuel Kant (2009) proposed some influential criticisms of it in *Perpetual Peace* (1795), and Albert Einstein, horrified by the nuclear culmination of World War II, declared that "a world government must be created which is able to solve conflicts between nations by judicial decision" (Einstein 2016).[17] Furthermore, Bostrom's "singleton hypothesis" posits that a world governing system constitutes the finalistic endpoint of geopolitical evolution (Bostrom 2006).[18] While there does appear to be some degree of historical momentum toward increasingly globalized forms of governance (the United Nations and European Union being the most salient examples), it is also true that "globalist trends" have been interrupted and reversed in the past—most notably with the failure of the League of Nations.[19] Nonetheless, a singleton appears to constitute the only plausible "world-configuration" that could effectively neutralize the global security threats looming in the twenty-first century. Without a global Leviathan to coordinate the actions of states, prevent interstate conflict, and

---

[15] Accountability also appears ineffective against suicidal attackers. In addition, metamaterial invisibility cloaks could further confound the transparent society model, as well as the ability for individuals to synthesize fake audio/video recordings that are virtually indistinguishable from real ones—a feat recently accomplished by researchers at the University of Washington. (See Suwajanakorn et al. 2017.)

[16] There are a couple of issues here that deserve further exploration. For example, omnipresent eyes will be unable to guarantee increased security if those they can't focus on the relevant risky phenomena. Furthermore, determining what the relevant risky phenomena are could render watching everyone unnecessary. Due to space constraints, I have bracketed such questions.

[17] See also Daniel Deudney's (2007) discussion of "nuclear one-worldism."

[18] Note the difference between "world government" and "world governance," the latter of which could be an intergovernmental body like the United Nations that, for example, coordinates state policies on a global level.

[19] Thanks to Matthijs Maas for clarification on this point.

neutralize agential risks, it is unclear how civilization can stave off an existential catastrophe, given Sotos' and my calculations presented in section 2.

Whereas section 2 discussed premise (i), this section has explored premises (ii) and (iii). Our tentative conclusion is that, following Persson and Savulescu, states will need to implement invasive surveillance systems to neutralize agential risks. But this may not be enough to avoid the Ultimate Harm of an existential catastrophe—rather, states themselves will need to coagulate under the aegis of a singleton. Yet, as we will see in the next section, which examines premises (iv) and (v), this proposal runs into several major problems.

4. Problems for Global Governance

Perhaps the most obvious objection to a global singleton is that "a single decision-making agency at the highest level" that includes among its powers "the ability to prevent any threats (internal or external) to its own existence and supremacy, and … to exert effective control over major features of its domain," to quote Bostrom (2006), has often resulted in bad outcomes for human well-being when implemented on the state-level. To be sure, an autocratic-like government run by truly *sagacious* leaders could be a force for good. There are regions—however small—of possibility space in which autocrats rule in benevolent ways, perhaps guided by a love of knowledge and wisdom, as with Plato's "philosopher king."[20] Yet there are inherent structural reasons why so few political rulers throughout history have been guided by such values.

For one, there are always ambitious rivals prepared to take one's place atop the pyramid of power. According to "selectorate theory," a crucial condition for avoiding this outcome is to control the flow of revenue and redistribute tax money to one's essential supporters; and naturally, the fewer of these supporters that one has, the better. In contrast, taking money from essential supporters and giving it to the poor would sour the alliances needed to maintain power, resulting in one being deposed or, worse, assassinated (see Bueno de Mesquita et al. 2003). Thus, however confident one might be that "If only *I* had all the power, I would—being a morally good person—change the world for the better," the nature of power structures forces people into less altruistic and more ruthless modes of political calculation. Even more, there is the obvious danger of megalomaniacal sociopaths ascending to the pinnacle of political preeminence. The result could be a totalitarian singleton that severely oppresses its population by exploiting an infrastructure of surveillance technologies that was, perhaps, put in place by prior, more benign regimes. This could yield an existential catastrophe like *permanent stagnation* (never reaching technological maturity) or *flawed realization* (reaching technological maturity but in a way that inevitably leads to subsequent failure) (see Bostrom 2013).

But even if a political leader heading a singleton were benevolent and incorruptible, the idea of creating a global singleton encounters an equally significant problem. To begin, consider that numerous justifications for the state and its "monopoly of the legitimate use of physical

---

[20] Note that a singleton could also take a democratic form. There are three reasons why I'm here focusing on autocracy: (i) space is limited, (ii) I follow Rawls (2002) in thinking that a singleton "would either be a global despotism or else would rule over a fragile empire torn by frequent civil strife as various regions and peoples tried to gain their political freedom and autonomy" (quoted at length in section 5), and (iii) the case ultimately considered in this paper involves a singleton run in autocratic fashion by a superintelligent machine.

force within a given territory," as Max Weber (1919) famously put it, have been proposed in the form of social contract theories. Those stemming from Thomas Hobbes' 1651 masterpiece, *Leviathan*, are generally termed "contractarian" while those associated with Kant's theories are "contractualist." For the present purposes, we will focus on the former. For Hobbes, the state's legitimacy derives from its capacity to provide security for its members in exchange for some degree of individual freedoms. The desideratum of security arises from a particular view of human nature, namely, that humans are both instrumentally rational and motivated by their own self-interests, according to psychological egoism.[21] Consequently, the "state of nature" is one marked by a war of all against all, in which life is "solitary, poor, nasty, brutish, and short." Incidentally, contemporary anthropological research suggests that Hobbes' hypothetical starting point, the state of nature, at least somewhat accurately reflects the general living conditions of early humans—i.e, the "original affluent society" paradigm that emerged in the 1960s and seemed to vindicate Jean-Jacque Rousseau's "noble savage" archetype is wrong.[22] Violence was much more prevalent back then than it is today and, according to Steven Pinker, the decline of violence in recent centuries is largely due to the rise of the Leviathan (Pinker 2011).

Insofar as one accepts this picture, casting one's eyes toward the future leaves a very worrisome afterimage. The reason is that the threat of universal unilateralism will not only increase the overall probability of doom but *also* undercut the social contract upon which a human-governed singleton would be jusficatorily founded. Consider that states are only able to satisfy their half of the social contract of providing security if and only if there exists a sufficiently large power differential between them and the citizens living within their borders. States not only act with legitimate force or violence, but just as crucially, they must have a *monopoly* on force or violence, so to speak. If the government of a state becomes unable to protect Joe from Sam and Sam from Bob, then the social contract will dissolve and, along with it, the modern state system.[23] As Wittes and Blum (2015) put this general point:

> This … remains an essential insight: we need the protection of a strong state as a precondition for the meaningful exercise of liberty. For this reason, the one common feature of … different contractarian visions is the promise … of the state to provide security, however defined, in exchange for the right to rule. … Technologies of mass empowerment threaten to undermine precisely this promise. Indeed, it is hard to contemplate a world in which anyone can attack anyone from anywhere, in which we have greatly distributed both the power and the vulnerability to attack, without thinking of Hobbes's state of nature, or what he called "warre," the situation from which the Leviathan state was meant to extricate us.

Let us call this the "threat of state dissolution":

---

[21] Although current neorealist traditions diverge from classical realism in identifying structural factors rather than human nature as responsible for the behaviors of states on the international level.
[22] Although Rousseau never used this term.
[23] Due to space limitations, I am here ignoring some alternative theses, such as Philip Bobbitt's notion of the "market state" (Bobbitt 2002).

Threat of state dissolution: a consequence of the democratization of dual-use technological capabilities is a reduction in the power differential between state and nonstate actors; extrapolating this forward, each is converging upon the same point of unprecedented power to obliterate civilization or cause human extinction (see Figure 2).[24]

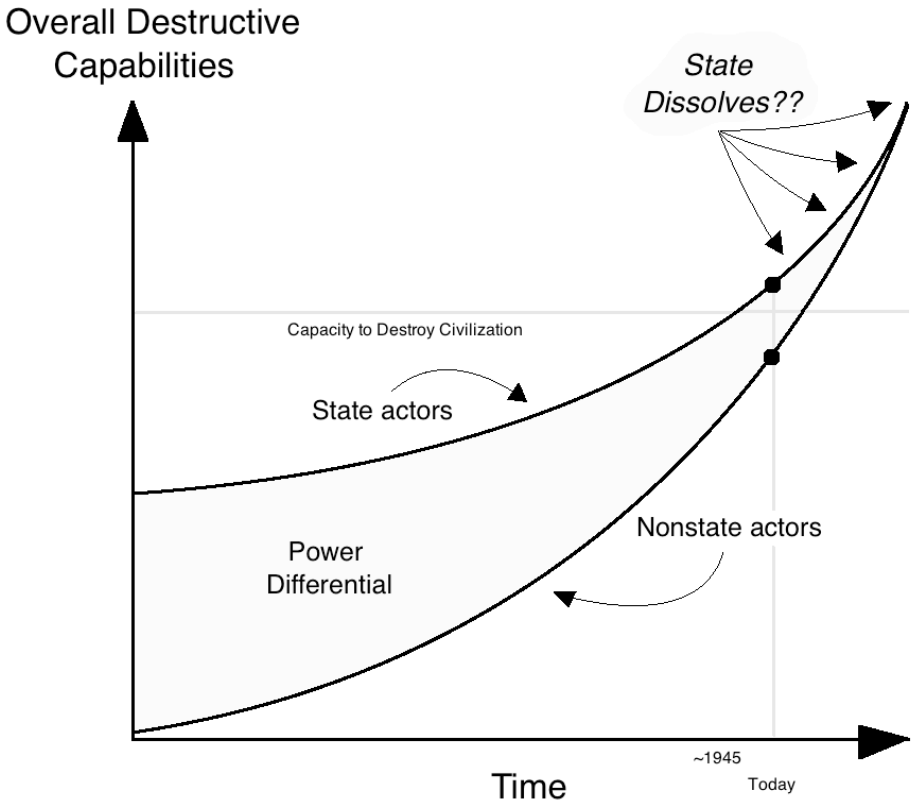

***Figure 2**: Schematic diagram of the convergence of state and nonstate power, ultimately leading to the dissolution of the modern state system.*

In fact, the relative reduction of traditional state power is already conspicuous in the contemporary world. Consider some concrete examples, beginning with the 9/11 attack perpetrated by al-Qaeda. This resulted in two major wars that could cost an estimated $6 trillion, left more than 8,000 coalition soldiers dead, and caused over 110,000 Iraqi civilian deaths. Yet, as Frances Flannery notes, circa 2001 "the core membership of Al Qaeda was most likely only around 500-1,000" (Flannery 2016).[25] Compare this to the US population at the time, which was 285 million, with ~1.3 million active duty military personnel. Or consider the 2016 Dyn cyberattack,

[24] Thus, what is not explicit here is that insofar as a singleton is a "state" in the relevant sense, which it is, then it too is subject to the vitiating phenomenon specified.
[25] To which Flannery adds that "unfortunately, a very small number of people can do an enormous amount of damage" (Flannery 2016).

which may have been perpetrated by a *single* "angry gamer" (Mathews 2016). Incredibly, the attack adversely affected a massive number of major websites, including Airbnb, Amazon, BBC, *The Boston Globe*, CNN, Comcast, *FiveThirtyEight*, Fox News, *The Guardian*, iHeartRadio, Imgur, National Hockey League, Netflix, *The New York Times*, PayPal, Pinterest, Pixlr, Reddit, SoundCloud, Squarespace, Spotify, Starbucks, Storify, the Swedish Government, Tumblr, Twitter, Verizon Communications, Visa, Vox Media, Walgreens, *The Wall Street Journal*, Wired, Yelp, and Zillow (to name a few). Or ponder the following hypothetical near-future scenario outlined by Stuart Russell:

> A very, very small quadcopter, one inch in diameter can carry a one- or two-gram shaped charge. You can order them from a drone manufacturer in China. You can program the code to say: "Here are thousands of photographs of the kinds of things I want to target." A one-gram shaped charge can punch a hole in nine millimeters of steel, so presumably you can also punch a hole in someone's head. You can fit about three million of those in a semi-tractor-trailer. You can drive up I-95 with three trucks and have 10 million weapons attacking New York City. They don't have to be very effective, only 5 or 10% of them have to find the target (quoted in Topol 2016).

This could be scaled up arbitrarily: perhaps a rogue state packs *100 million* of these weapons into *hundreds* of semi-trucks around the world and then deploys this drone army within a five-minute window. The consequences could be as severe as a nuclear war or global pandemic (Torres 2017a). Yet the weaponization of modified drones could fall within the sphere of feasibility for nonstate agential risks as well. Thus, Russell notes that "there will be manufacturers producing millions of these weapons that people will be able to buy just like you can buy guns now, except millions of guns don't matter unless you have a million soldiers. You need only three guys to write the program and launch [these drones]" (Topol 2016).

It follows that the threat of universal unilateralism might very well preclude the only available option that could save humanity from the existential dangers posed by the threat of universal unilateralism. Without a monopoly on force, violence, and power, the singleton that we may *need* to ensure our survival on spaceship Earth will stand no taller, so to speak, than the groups and individuals that reside within its terrestrial domain.[26] This leaves us with the unsavory conclusion that an existential catastrophe in the coming decades or centuries could be extremely probable (see here Torres 2017e).

5. The Friendly Supersingleton Hypothesis

So far we have considered human institutions run by human beings to ensure the security of human beings. With respect to the actors involved, the playing field is fundamentally level. But what happens when a friendly superintelligence enters the picture? By "friendly," I mean the value system that defines its utility function is sufficiently aligned with our "human

---

[26] One might here suggest that we should expand into space. But for the many reasons outlined in Torres 2017f, the colonization of space could have truly catastrophic consequences, resulting in an s-risk or "suffering risk."

values" (whatever they are) to ensure a good outcome for our species; in other words, its relation with humanity is marked by *amity* rather than *enmity*. This section will argue that a friendly superintelligence at the helm of a global singleton could provide an escape from the labyrinthine catch-22 outlined by the previous sections.

To begin, whereas it would be unwarranted to assert that intelligence *is* power, it would not be unwarranted to assert that intelligence *yields* power. Our own species provides an example: the unparalleled dominance of *Homo sapiens* within the Animal Kingdom stems not from our sharp teeth, quick speed, long claws, opposable thumbs, or bipedal posture—although the latter two have been instrumentally useful and may have coevolved in crucial ways with the rise of our encephalization quotient (EQ). Rather, it is our superior intelligence—or problem-solving capacity—that has enabled us to subjugate a large portion of the Gaian system for our own personal benefit. Thus, a computer program with greater intelligence than what is attainable in principle by any organism with a human-specific genome would find itself able to control the physical world in even more profound ways (see Bostrom 2014; Yampolskiy 2016).[27]

There are two properties in particular that could bestow immensely greater power to a superintelligence. The first is *quantitative*: the information processing abilities of silicon (or carbon nanotube) hardware exceed those of our neural wetware by orders of magnitude. More specifically, computers can process information about one million times faster than our brains, meaning that a single minute of objective time would equal about 2 years of subjective time for an uploaded mind. From its perspective, the outside world would be virtually frozen in place. This could enable it to solve a wide variety of complex problems on timescales that could appear almost instantaneous to us. Whereas it takes the average PhD student 8.2 years or so to attain the highest level of expertise on a specialized topic, this could be achieved by a quantitative superintelligence in a matter of 4.3 minutes.

Even more, whereas human brains have a storage capacity of between 10 and 100 terabytes, a superintelligence's "memory" would be limited only by the hardware available to it—and the available hardware will likely be extensive by the time the first superintelligence is created or creates itself (that is, through recursive self-improvement). Indeed, a superintelligence connected to the Internet could make immediate use of this network as its "extended mind," much the same way that Wikipedia constitutes a kind of neuroprosthesis for many humans today, storing vast amounts of data so that our hippocampuses (and other brain structures) don't have to (see Clark and Chalmers 1998). Whereas collective human knowledge has grown exponentially since the Scientific Revolution, the human brain has remained more or less fixed and finite. The result is an exponential growth of *relative individual ignorance*, measured as the difference between the total knowledge had by the collective whole and the average individual. A superintelligence, however, could rectify this situation for *itself* by making everything known something that it knows. But of course it could also generate its own knowledge about the world—and at a pace that not even the collective intellect of humanity could keep up with. In sum, the superhuman abilities to process and retain information would give a quantitative superintelligence an im-

---

[27] That is, the most formidable long-term threat to human survival *that we know of*. There could be any number of more serious future risks currently hidden beneath the horizon of our collective imagination.

mense strategic advantage over humanity—an advantage that it could use for good, if friendly, or ill, if unfriendly.

The second property is *qualitative*: the "concept-generating mechanisms" of a superintelligent mind could be different in kind from those lodged in our biological brains. The idea is this: concepts are mentalistic entities that *represent* some feature of mind-independent reality, including processes and objects. We thus have concrete concepts for chairs and automobiles, abstract concepts for democracy and justice, and processual concepts for running and jumping. If our brains were unable to generate any of these concepts, our minds would be unable to represent the corresponding features of reality, resulting in something akin to a *cognitive scotoma* (or mental "blind spot"). This being said, the concept-generating mechanisms that we *do* have were given to us by contingent evolution; it follows that a species' evolutionary history will determine the circumscribed range of concepts that it can generate—call the resulting territory of knowability its "cognitive space." Whereas humans can generate the concepts of, for example, *nuclear chain reaction* and *big bang*, these most certainly fall outside the cognitive space of chipmunks and grasshoppers, as well as chimpanzees and bonobos. For these creatures, it is not merely a matter of lacking the relevant knowledge, but of being unable to ever acquire that knowledge *in principle*.

While a whole brain emulation (or mind-upload) would inherit the concept-generating mechanisms had by the human brain of which it is a clone, I would conjecture—following Bostrom (2014) and others—that neuromorphic or directly programmed AI are more likely to emerge as the first greater-than-human-level intelligences. Due to limitations on space, I won't justify this claim. Suffice it to say that both could instantiate radically alien cognitive architectures that correspond to cognitive spaces that subsume and/or far exceed our own cognitive spaces. The result would be access to concepts that lie forever beyond our epistemic reach—and with a new library of concepts, a qualitative superintelligence could represent reality in completely novel ways. For example, it could identify features of the universe that enable it to construct entirely new causal theories and perhaps even an entirely new *physics*. It could then use these theories to invent novel ways of manipulating the physical world that would eternally perplex the human mind, much the same way that computers, jet planes, space travel, cell phones, and the like are eternally perplexing to chipmunks (insofar as chipmunks can even *be* perplexed by such technological "magic"). The point is that the ability to *make things happen* in the universe by pulling levers and wiggling mechanisms hidden behind the curtain of human comprehension would also bestow an immense strategic advantage over humanity (Torres 2017a).[28] (See Figure 3.)

This brings us to the governing issues outlined above. Put simplistically: since intelligence yields power, a superintelligence would be superpowerful. Its relationship with humanity would be more akin to the *interspecies* dominance of humans over gorillas than the *intraspecies* dominance of, say, a CEO over her employees or the United States over a country like Luxembourg. In other words, the vertical power differential between us and it could be quite vast, with it possessing the kind of monopoly on force and domination that characterizes our relations with

[28] Note that while a superintelligence could arise with quantitative but not qualitative characteristics, it is unlikely to arise with qualitative but not quantitative characteristics. Thus, we should expect either a quantitative-only superintelligence or a quantitative-qualitative superintelligence to arise, if one does.

other, "lower" species on the planet, many of whose fate now depends upon the wisdom and benevolence of our collective decision-making. This suggests that a superintelligent machine could potentially re-establish a social contract—call it a *post-singularity social contract*—whereby all humans give up the right to govern in exchange for security against the growing threat of universal unilateralism, given the superintelligence's capacity to overcome the growing threat of state dissolution. This contract could thus form the justificatory basis of a global "super-singleton" that could protect humanity from a wide range of possible harms, including, at the extreme, existential risks. It could accomplish this end by using the aforementioned strategies of information collecting and social control, as well as some anti-risk enforcement program not yet imagined (or even imaginable by the human mind).

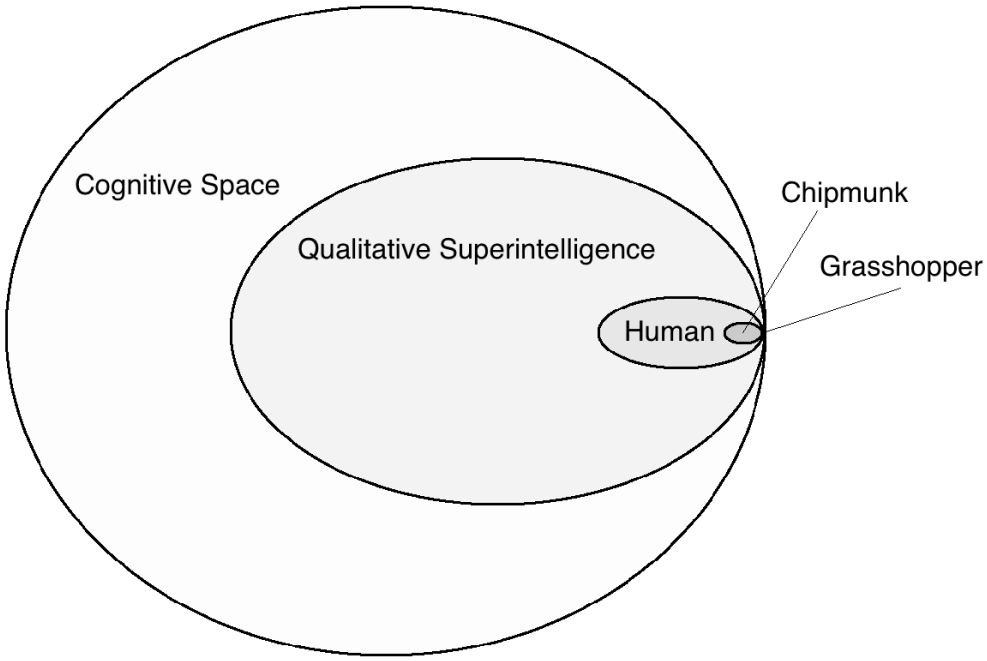

Figure 3: *Schematic illustration of cognitive space (Torres 2017a)*

For this radical proposal to work, of course, the superintelligence—perhaps designed specifically for the purpose of governing, perhaps with a "super-persuader" capacity that would make physical force unnecessary (Bostrom et al. 2017)[29]—would need to be *friendly*, as defined at the beginning of this section. Yet the control problem, i.e., the challenge of ensuring that a superintelligence is friendly, appears to be one of the most formidable, high-stakes problems that

[29] Note that a "super-persuader" superintelligence might not need *force* to prevent, say, omnicidal agents from annihilating humanity. It could, instead, simply talk such individuals out of causing harm. See Bostrom et al. 2017.

humanity has ever had to confront. Given that there are far more ways to get the control problem wrong than right, my own view is that successfully creating a friendly superintelligence is, all things considered, less probable than screwing things up, perhaps irreversibly (see Torres 2017e). Yet if we *do* get the control problem right, the outcome could be not merely *good* but genuinely *utopian*. Consider Bostrom's (2009) claim that

> one might believe that superintelligence will be developed within a few centuries, and that, while the creation of superintelligence will pose grave risks, once that creation and its immediate aftermath have been survived, the new civilization would have vastly improved survival prospects since it would be guided by superintelligent foresight and planning.

By controlling the global economy, repairing the environment, eliminating interstate arms races and wars, and neutralizing the threat posed by agential risks, the probability of an existential catastrophe could fall below the historical level of our *cosmic risk background*. (Indeed, advanced technologies could also be used to overcome threats from nature, such as asteroid impacts and supervolcanic eruptions.) Even more, if some moral "ought" statements can be reduced to descriptive "is" statements, a superintelligence could use the tools of science to devise legal norms that maximally enhance the human (or posthuman) condition (see Harris 2010). For example, once one accepts the moral prescription to maximize human well-being, it becomes a merely *empirical question* how best to achieve this. One can thus conduct experiments (perhaps in the form of simulations) to see whether, say, free market systems produce more human well-being than democratic socialist systems, religion produces more human well-being than atheism, or psychodynamic therapy is more effective at overcoming mental illnesses than cognitive behavioral therapy. Since an *instrumental* value of superintelligence is likely to be the acquisition of a complete "theory of everything" (because this would facilitate a wide range of final goals), a superintelligence could obtain extensive knowledge about which social, cultural, political, economic, and so on configurations are most conducive to human prosperity. The result could be something like the "best of all possible worlds": a system designed to make unhappy people happy and happy people even happier.

This proposal also circumvents many of the concerns that scholars have articulated in the context of theorizing about world governing systems. For example, Kant argued against the idea of global governance because, he claimed, such a system would be *ineffective* at enforcing law and order. Perhaps this is true in the case of human leadership (I am in fact inclined to agree), but for reasons discussed above, it is not a compelling objection with respect to *superhuman* leadership. Along these lines, John Rawls (2002) writes,

> I follow Kant's lead in *Perpetual Peace* in thinking that a world government—by which I mean a unified political regime with the legal powers normally exercised by central governments—would either be a global despotism or else would rule over a fragile empire torn by frequent civil strife as various regions and peoples tried to gain their political freedom and autonomy.

Having already addressed the second disjunct (i.e., a superintelligence-controlled singleton wouldn't be ineffective), consider the first: it is true that a superintelligence at the helm of a singleton, as here envisaged, would be something like a despot or dictator. But here—in this very specific context—we need to divest these terms of their negative connotations and try to glimpse this situation from a radically different Gestalt. Whereas human beings are myopic, foolish, venal, and self-serving, a friendly superintelligence wouldn't embody any of these negative characteristics *by definition*. Rather, it would rule as a benevolent hegemon, considering the opinions and preferences expressed by individuals under its aegis, but ultimately making policy decisions based on its own judgments, founded on the various values—e.g., human security, prosperity, liberty, freedom, and universal rights[30]—that its programmers loaded into it. True, society would become a little "less liberal" in a sense, yet losing certain freedoms to a value-aligned superintelligent machine could entail more total freedom than ever before within the lower-level realm of human affairs.

There are a few important conclusions that emerge from this discussion. First, everything hangs on our ability to solve the control problem and create a friendly superintelligence capable of wise governance. This challenge is formidable enough given that many AI experts anticipate a human-level AI within this century (Müller and Bostrom 2014)—meaning that there appears to be a deadline—but the trends outlined in Figure 2 open up the possibility that we may have *even less time* to figure out what our "human values" are and how they can be encoded in "the AI's programming language, and ultimately in primitives such as mathematical operators and addresses pointing to the contents of individual memory registers" (Bostrom 2014). Thus, the present paper offers a novel reason for allocating large amounts of resources for projects that focus on solving the control problem: not only will continued progress in computer science make the control problem probably unavoidable, but the convergence of state and nonstate power could require new forms of global governance—namely, a friendly supersingleton—within the coming decades.[31]

There is yet another way to look at this proposal. One could object that the idea of a superintelligence controlling a global regime is outrageous and crazy. It is a fantasy because "Friendly AI" is nothing more than pure magic.[32] To this one could respond, somewhat sardonically, that as Arthur Clarke's third law states, "any sufficiently advanced technology is indistinguishable from magic." Thus, however "magical" a friendly superintelligence may seem to our limited minds at the mid-morning of the twenty-first century—only slightly more than 80 years after the first electro-mechanical binary programmable computer was invented—this does not constitute a cogent reason for rejecting the above arguments. The more forceful response is to

---

[30] For an argument against the social choice ethics approach, see Baum forthcoming.

[31] One issue not discussed is how probable the rise of superintelligence is. First, I believe that intelligence, understood in philosophical terms as equivalent to *instrumental rationality*, is algorithmic in nature and can be multiply instantiated in any physical system that exhibits the right functional organization. Second, recent surveys of AI experts suggest that the probability of a superintelligence joining humanity on our pale blue dot before the year 2100 is nearly 100 percent (Müller and Bostrom 2014; see also Sotala and Yampolskiy 2014). Put differently, there is an extremely good chance—if the experts are to be believed—that a child born today will live to witness the rise of machine superintelligence. This further supports the claim that, if theorists managed to solve the control problem and computer scientists manage to successfully program human values into the AI, the ideas presented here are *realistic*.

[32] This response to some actual criticisms of this paper.

say this: "Well, then, how do *you* propose that humanity survives the dual threats of universal unilateralism and state dissolution?" To quote Wittes and Blum (2015) once more,

> if technologies of mass empowerment enable many isolated individuals or diverse non-state actors to injure or violate other individuals on a mass scale anywhere around the globe with substantially reduced fear of detection and punishment, [then] *we are in big trouble*. Much of civil and political life as we know it will likely come to an end.[33]

Unless critics can propose a good case for rejecting the calculations that lead Sotos (2017) and myself (2017e) to hypothesize a Great Filter up ahead, we will need to invent *some global-scale macro-strategy* for preemptively neutralizing state and nonstate actors from blowing up the world with dual-use emerging technologies, whether by error or terror. Making matters worse, there are also reasons for believing that expanding into space is not a promising solution to this problem: as I elsewhere show, space colonization will almost certainly yield constant, devastating wars between planetary civilizations that result in astronomically huge amounts of suffering, i.e., an "s-catastrophe" (Torres 2017f; see also Deudney, forthcoming). There is, as some environmentalists say, no "Planet B" to seek refuge on. Humanity should want to remain on Earth, but remaining on Earth will require that we address the phenomena of premises (i) through (v) in section 1.

The present discussion also bears on a question sometimes propounded in debates about AI risk: "If superintelligence poses an existential risk to humanity, then why not abandon research on the topic? Why not relinquish this line of research?" The first rejoinder is that this appears *infeasible* due to Winner's autonomous technology thesis and Bostrom's technological completion conjecture. The second rejoinder is that relinquishing this technology appears *undesirable* given the threats of universal unilateralism and state dissolution. At least by creating a superintelligence—especially one specifically designed to be a "super-governor"—we stand a chance of surviving the democratization of science and technology.[34]

6. Conclusion

Michael Walzer (2004) once declared that "the dream of a single agent—the enlightened despot, the civilizing imperium, the communist vanguard, the global state—is a delusion." This might be true within the paradigm of human leadership, but it is probably false within the paradigm of a superintelligent regime. An "enlightened despot" in the form of a superintelligence ruling the world as a friendly supersingleton could usher in a new age of peace and prosperity. It could constitute what some call an *existential eucatastrophe*, or "an event which causes there to be much more expected value after the event than before" (Cotton-Barratt and Ord 2015). Even more, without such a system in place, the democratization of science and technology could all but guarantee an existential catastrophe. To borrow an aphorism from Voltaire, "*Si Dieu n'existait pas, il faudrait l'inventer*," meaning, "If God did not exist, it would be necessary to invent

---

[33] Italics added.
[34] Thanks to Matthijs Maas for pointing this out to me.

him." Given the global security predicament of tomorrow, the present chapter agrees—that is, if "God" takes the form of a value-aligned superintelligent machine.[35]

Acknowledgments

Thanks to Kaj Sotala, Stuart Armstrong, Catherine Rhodes, Anders Sandberg, Olle Haggstrom, Karin Kuhlemann, Karim Jabari, Markus Stoor, James Miller, Markus Salmela, Timoteus Dahlberg, Alexey Turchin, and all other participants of the 2017 Existential Risks to Humanity workshop at Chalmers University of Technology, September to October 2017. Thanks also to an anonymous referee for helpful comments. Special thanks to Matthijs Maas, Kenny Easwaran, and Justin Shovelain for extensive, insightful feedback.

References

Allison, Graham. 2004. *Nuclear Terrorism: The Ultimate Preventable Catastrophe*. New York, NY: Henry Holt and Company, LLC.

Baum, Seth. 2014. The Great Downside Dilemma for Risky Emerging Technologies. *Physica Scripta*. 89(12). http://iopscience.iop.org/article/10.1088/0031-8949/89/12/128004.

Baum, Seth. forthcoming. Social Choice Ethics in Artificial Intelligence. *AI & Society*. http://www.sethbaum.com/ac/fc_SocialChoice.pdf.

Baum, Seth, Timothy Maher, and Jacob Haqq-Misra. 2013. Double Catastrophe: Intermittent Stratospheric Geoengineering Induced by Societal Collapse. *Environment Systems & Decisions*. 33(1): 168-180.

Bobbitt, Philip. 2002. *The Shield of Achilles: War, Peace, and the Course of History*. New York, NY: Knopf.

Boles, Kent, Krishna Kannan, John Gill, Martina Felderman, Heather Gouvis, Bolyn Hubby, Kurt Kamrud, J. Craig Venter, and Daniel Gibson. 2017. Digital-to-biological converter for on-demand production of biologics. *Nature Biotechnology*. 35: 672-675.

Bostrom, Nick. 2002. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*. 9(1).

Bostrom, Nick. 2005. Transhumanist Values. *Review of Contemporary Philosophy*. 4.

Bostrom, Nick. 2006. What is a Singleton? *Linguistic and Philosophical Investigations*. 5(2): 48-54.

---

[35] Note that parts of this chapter draw from Torres 2016 and 2017a, in some cases *ad verbum*.

Bostrom, Nick. 2009. The Future of Humanity. In *New Waves in Philosophy of Technology*, edited by Jan-Kyrre Berg Olsen, Evan Selinger, and Soren Riis. New York, NY: Palgrave McMillan.

Bostrom, Nick. 2013. Existential Risk Prevention as Global Priority. *Global Policy*. 4(1): 15-31.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford University Press.

Bostrom, Nick, Thomas Douglas, and Anders Sandberg. 2016. The Unilateralist's Curse: The Case for a Principle of Conformity. *Social Epistemology*. 30(4): 350-371.

Bostrom, Nick, Allan Dafoe, and Carrick Flynn. 2017. Policy Desiderata in the Development of Superintelligent AI. Working Draft. https://nickbostrom.com/papers/aipolicy.pdf.

Brin, David. 1996. The Transparent Society. *Wired*. Accessed on July 27, 2017. https://www.wired.com/1996/12/fransparent/.

Bueno de Mesquita, Bruce, Alastair Smith, Randolph M. Siverson, and James Morrow. 2003. *The Logic of Political Survival*. Cambridge, MA: MIT Press.

Clark, Andy, and David Chalmers. 1998. The Extended Mind. *Analysis*. 58: 10-23.

Cotton-Barratt, Owen, and Toby Ord. 2015."Existential Risk and Existential Hope: Definitions. FHI Technical Report. Accessed on July 27, 2017. https://www.fhi.ox.ac.uk/Existential-risk-and-existential-hope.pdf.

Deudney, Daniel. 2007. *Bounding Power: Republican Security Theory from the Polis to the Global Village*. Princeton, NJ: Princeton University Press.

Diamond, Jared. 1999. The Worst Mistake in the History of the Human Race. *Discover*. Accessed on July 27, 2017. http://discovermagazine.com/1987/may/02-the-worst-mistake-in-the-history-of-the-human-race.

Drexler, Eric. 2013. *Radical Abundance: How a Revolution in Nanotechnology Will Change Civilization*. New York, NY: PublicAffairs.

Dunbar, Robin. 1998. *Grooming, Gossip, and the Evolution of Language*. Cambridge, Mass.: Harvard University Press.

Einstein, Albert. 2016. *The Albert Einstein Collection: Essays in Humanism, The Theory of Relativity, and The World As I See It*. Philosophical Library/Open Road.

Flannery, Frances. 2016. *Understanding Apocalyptic Terrorism: Countering the Radical Mindset*. New York, NY: Routledge.

Forge, John. 2010. A Note on the Definition of "Dual Use." *Science and Engineering Ethics*. 16(1): 111-118.

Fukuyama, Francis. 1992. *The End of History and the Last Man*. New York, NY: Avon Books, Inc.

Fukuyama, Francis. 2002. Our Posthuman Future: Consequences of the Biotechnology Revolution. New York, NY: Farrar, Straus, and Giroux.

Hanson, Robin. 2003. Shall We Vote on Values, but Bet on Beliefs? http://mason.gmu.edu/~rhanson/futarchy2013.pdf.

Harris, Sam. 2010. *The Moral Landscape: How Science Can Determine Human Values*. New York, NY: Free Press.

Hobbes, Thomas. 1651 (2016). *Leviathan or The Matter, Forme and Power of a Common Wealth Ecclesiasticall and Civil*. CreateSpace Independent Publishing Platform.

Huntington, Samuel. 1993. The Clash of Civilizations? *Foreign Affairs*. 72(3): 22-49.

Kant, Immanuel. 2009. Kant's Vision of a Just World Order. In *The Blackwell Guide to Kant's Ethics*, edited by T.E. Hill. New York, NY: Blackwell, 196–208

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York, NY: Viking.

Langman, Peter. 2010. *Why Kids Kill: Inside the Minds of School Shooters*. New York, NY: St. Martin's Griffin.

Mann, Steve, Jason Nolan, and Barry Wellman. 2003. Sousveillance: Inventing and Using Wearable Computing Devices for Data Collection in Surveillance Environments. *Surveillance and Society*. 1(3): 331–355

Mathews, Lee. 2016. Angry Gamer Blamed for Most Devastating DDoS of 2016. *Forbes*. https://www.forbes.com/forbes/welcome/?toURL=https://www.forbes.com/sites/leemathews/2016/11/17/angry-gamer-blamed-for-most-devastating-ddos-of-2016/&refURL=https://en.wikipedia.org/&referrer=https://en.wikipedia.org/#78871c472dac.

Miller, Jeff, and Wolf Schwarz. 2014. Brain signals Do Not Demonstrate Unconscious Decision Making: An Interpretation Based on Graded Conscious Awareness. *Consciousness and Cognition*. 24: 12-21

MIT. 2017. Genetic Engineering Holds the Power to Save Humanity or Kill It. *MIT Technology Review*. https://www.technologyreview.com/s/608903/genetic-engineering-holds-the-power-to-save-humanity-or-kill-it/.

Mukunda, Guatam, Kenneth Oye, and Scott Mohr. 2009. What Rough Beast? Synthetic Biology, Uncertainty, and the Future of Biosecurity. *Politics and the Life Sciences*. 28(2): 2-26.

Müller, Vincent, and Nick Bostrom. 2014. Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In *Fundamental Issues of Artificial Intelligence*, edited by Vincent Müller. Berlin, Germany: Springer.

Persson, Ingmar, and Julian Savulescu. 2012. *Unfit for the Future: The Need for Moral Enhancement*. Oxford, UK: Oxford University Press.

Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. New York, NY: Viking.

Rawls, John. 2002. *The Law of Peoples: With, The Idea of Public Reason Revisited*. Cambridge, MA: Harvard University Press.

Robock, Alan, Luke Oman, Georgiy Stenchikov, Owen Toon, Charles Bardeen, and Richard Truco. 2007. Climatic Consequences of Regional Nuclear Conflicts. *Atmospheric Chemistry and Physics*. 7(8): 2003-2012.

Soon, Chun Siong, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes. 2008. Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*. 11 (5): 543-5.

Sotala, Kaj, and Roman Yampolskiy. 2014. Responses to Catastrophic AGI Risk: A Survey. *Physica Scripta*. 90(6).

Smith, Kerri. 2008. Brain Makes Decisions Before You Even Know It. *Nature*. Accessed on July 27, 2017. http://www.nature.com/news/2008/080411/full/news.2008.751.html.

Sotos, John. 2017 Biotechnology and the Lifetime of Technical Civilizations. arXiv.org. https://arxiv.org/abs/1709.01149.

Suwajanakorn, Supasorn, Steven Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: Learning Lip Sync from Audio. *ACM Transactions on Graphics*. 36(4).

Topol, Sarah. 2016. Attack of the Killer Robots. *BuzzFeed*. https://www.buzzfeed.com/sarahatopol/how-to-save-mankind-from-the-new-breed-of-killer-robots.

Torres, Phil. 2016. *The End: What Science and Religion Tell Us About the Apocalypse*. Durham, NC: Pitchstone Publishing.

Torres, Phil. 2017a. *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks*. Durham, NC: Pitchstone Publishing.

Torres, Phil. 2017b. Agential Risks and Information Hazards: An Unavoidable but Dangerous Topic? Forthcoming in *Futures*.

Torres, Phil. 2017c. Who Would Destroy the World? Omnicidal Agents and Related Phenomena. Forthcoming in *Aggression and Violent Behavior*.

Torres, Phil. 2017d. Moral Bioenhancement and Agential Risks: Good and Bad Outcomes. *Bioethics*. 31:691-696.

Torres, Phil. 2017e. Facing Disaster: The Great Challenges Framework. Forthcoming in *Foresight*.

Torres, Phil. 2017f. Space Colonization and Existential Risks: On Why Following the Maxipok Rule Could have Catastrophic Consequences. Working draft: https://goo.gl/rvDdLj.

Walzer, Michael. 2004. *Arguing about War*. New Haven, CT: Yale University Press.

Weber, Max. 1919 (2004). *The Vocation Lectures*. Indianapolis, IN: Hackett Publishing Company, Inc.

Winner, Langdon. 1977. *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought*. Cambridge, MA: MIT Press.

Wittes, Benjamin, and Gabriella Blum. 2015. *The Future of Violence: Robots and Germs, Hackers and Drones—Confronting a New Age of Treat*. New York, NY: Basic Books.

Yampolskiy, Roman. 2016. *Artificial Superintelligence: A Futuristic Approach*. Boca Raton, FL: Taylor & Francis Group.

Yudkowsky, Eliezer. 2008. Artificial Intelligence as a Positive and Negative Factor in Global Risk. Machine Intelligence Research Institute. https://intelligence.org/files/AIPosNegFactor.pdf.